Introduction to Data Analysis

Elaborated in the framework of

SUB-FLY Project

financed under the contract

SEE EY-COP-0082

by

SEE Grants 2014-2021

Working together for a green, competitive and inclusive Europe

SEE & Cha(lle)nge

Sustainable & applied education

This material is elaborated as support for all students, staff, professionals and entities interested to analyze the status of a company using Business Canvas Model. It is used as a training support for students involved in *SUB-FLY Project.*

We are thankful to **SEE Grants 2014-2021**, which financed the *SUB-FLY Project* under the contract EY-COP-0082. Faculty of Business from Babes-Bolyai University (Romania) and School of Business from University of South-Eastern Norway (Norway) received, based on this contract, the opportunity to cooperate. One of the outcomes generated by this cooperation consists in development and implementation of the *SUB-FLY Project*.

This document was realized with the EEA Financial Mechanism 2014-2021 financial support. Its content (text, photos, videos) does not reflect the official opinion of the Programme Operator, the National Contact Point and the Financial Mechanism Office. Responsibility for the information and views expressed therein lies entirely with the author(s).

Introduction to Data Analysis

1. Preliminary notes

1.1 Brief history

Basic elements of statistics have been used since the beginning of civilization.

During Peloponnesian war (431 – 404 B.C), Athenians defending Plataea have counted the layers of bricks used to build the walls of Spartan fortification surrounding Plataea. Several soldiers were counting, and the most frequent value was considered to be the correct one. Thus, the Athenian defenders estimate the height of walls and escape the Spartan led siege. This technique is the currently well-known concept of mode.

The Han Dynasty and the Roman Empire were among the first states collecting data on the size of population, geography, and wealth of the empire.

Al-Kindi (801-873 AD) used for the first-time analysis of frequencies to break ciphers.

In 1662, John Graunt and William Petty have created the first life table, computing the survival probability for each age group, and thus they estimated London's population.

Starting with 16th century, development of probability theory is placing statistics on a new path, as a branch of mathematics. Gottfried Achenwall is using in 1749 the term *STATISTIK* to describe analysis of data about the state. The concept of STATISTIK was introduced into English in 1791, by Sir John Sinclair while publishing *Statistical Account of Scotland.*

Pierre de Fermat, Blaise Pascal (probability theory), Jakob Bernoulli, Abraham de Moivre (mathematical bases of statistics), Thomas Bayes (Bayes formula for computing probability), Pierre-Simon Laplace, Carl Friedrich Gauss (normal distribution), William Playfair (charts), Antoine Augustine Cournot (median), Karl Pearson (first Department of Statistics in a university – University College London, chi-square test, correlation), Ronald Fisher (ANOVA test), William Sealy Gosset (Student distribution), Jerzy Neyman (confidence intervals) are among mathematicians and scientists contributing to the development of statistics.

1.2 Definition of statistics

Initial purpose of statistics was to provide demographic data to state authorities. Starting with the 19th century, statistics becomes a branch of mathematics, and the concept is extended and applied in several fields of science and economy.

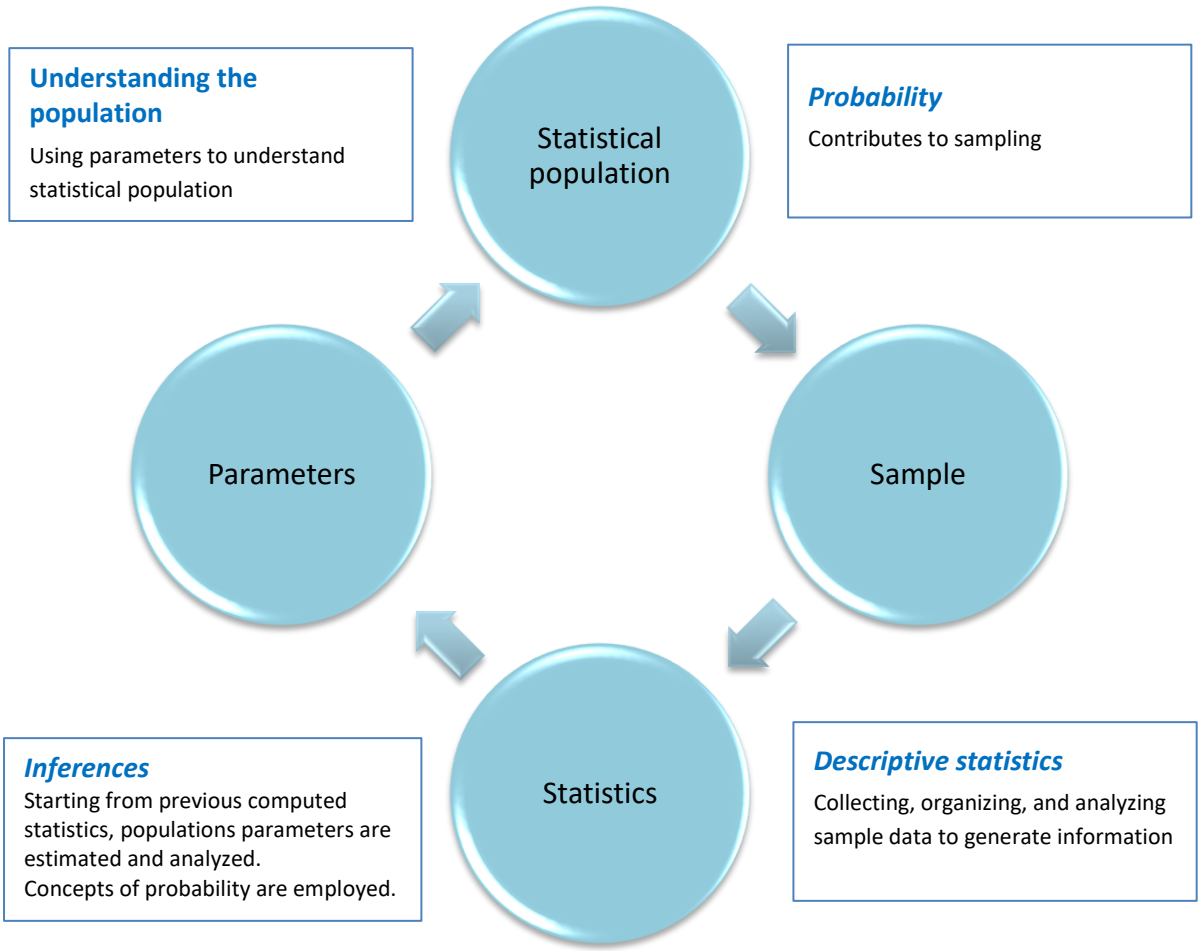
Statistics means using data, in an uncertain context, to generate information. Thus, the fundament for decision making is created.

In current society, our actions are based on data and generate new data. Analysing these data is based on specific techniques of statistics.

Thus, statistics might be defined as the science of collecting and analysing data for generating information about a population or phenomenon.

1.3 Concepts

Statistics is studying an entire population and not only single elements of the population. Thus, information provided have a certain confidence level, which can never be 100%.



Well-known and widely used parameters and their corresponding sample statistics are average, mode, median, proportion.

There are two types of studies:

1. **Sample survey –** elements of a population are analyzed to determine some characteristics of the entire population.
2. **Comparative survey** – elements from a population are compared with other elements from the same population or from a different one to determine some differences or to study a phenomenon.

We briefly present some of the basic concepts required for data analysis. Suppose we intend to perform a study for all companies registered in Cluj County.

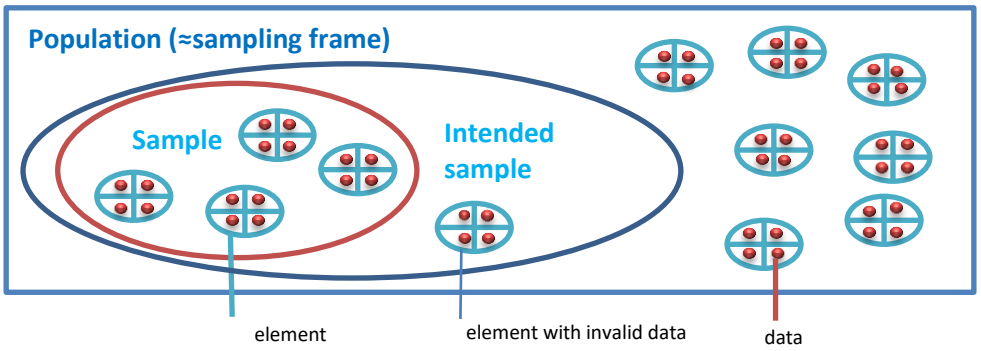
1. **Elements** are the entities, persons or objects owning data (information).
   Example: a company registered in Cluj County.
   A student, an employee, a customer might be other examples of elements.
2. **Population** is the set of all elements. It's a general collection.
   Example: all companies registered in Cluj County.
   All customers of Dell, all employees of Sperre, all students of BBU might be other examples of population.
   Number of all elements from a population is referred as volume. In general populations have large volumes, thus being inefficient to analyze the entire population, studies being performed on subsets of the population. One exception is a census.
   Volume of a population is denoted by N.
3. **Sampling frame** is a list of all elements. It's a specific collection.
   Example: list of names for all companies registered in Cluj County: For Your Team, Kudos Technologies, Fivetech Software Solutions, Vest TransCom, Magic Tour, Magic Fashion, Schuller, …
   Ideally, the sampling frame should cover the entire population. Due to the large volume of the population this is not always feasible.
   It is mandatory that the sampling frame is representative for the population. This means that sampling frame covers all groups of the population. For example, when analyzing the population of all companies registered in Cluj County, if transportation companies are skipped from the sampling frame, then the sampling frame is not representative for the population.
4. **Intended sample** – is the subset of the population from which we intend to collect data. Not all elements targeted in the intended sample will provide data, some of the reasons being: refuse to join the study, incomplete or invalid answers. Intended sample must contain enough elements to ensure that the required data are collected.
   Example: 20% off all companies registered in Cluj County. We must know exactly which are these companies and thus we might consider the following: For Your Team, Kudos Technologies, Fivetech Software Solutions, Vest TransCom.
5. **Sample** – it the subset of the population providing data used in the study. Sample must be representative for the population, meaning that it must have a certain structure and a certain volume.
   Example: 18% off all companies registered in Cluj County. 2% (difference between sample and intended sample) are the companies refusing to join the study, missing contact details

or providing incomplete or invalid data. We might get data from the following companies: For Your Team, Kudos Technologies, Fivetech Software Solutions.

Volume of a sample is denoted by n.

6. **Data** – information provided by each element of the sample and used in the study.
   When expressing a certain characteristic, data are referred as ***qualitative.***
   Example: town of residence, degree of satisfaction, field of activity, year of registration, number of stars for a hotel.
   When expressing discrete numbers (integers but not decimal numbers), data are referred as **quantitative discrete.**
   Example: number of employees, number of active customers, number of products in the portfolio, number of defective products per month.
   When expressing continuous numbers (decimal numbers are accepted), data are referred as **quantitative continuous.**
   Example: turnover, profit, efficiency rate, defective rate, height, temperature, weight.
7. **Variable** is a property of data used to organize and analyze them.

The following diagram is synthesizing the basic concepts presented above:



2. *Data collection*

Among purposes of data analysis are to offer a helicopter view for a phenomenon or substantiation of a managerial decision. First step it the study is to clarify what we intend to study (question(s) which will be answered at the end of our study). Subject of the study is influencing how we collect data, which statistical tools we might employ during the analysis and what type of data (qualitative, quantitative discrete or quantitative continuous) we need.

Main sources of data are databases and questionnaires.

2.1 Collecting data from data bases

Databases are an efficient (easy, quick, and affordable) tool to collect data. Structure and type of data, predefined by the database represent the main disadvantage.

Some useful databases are: National Institute of Statistics, Eurostat (https://ec.europa.eu/eurostat/web/main/home), World Bank (https://data.worldbank.org/), OECD (https://data.oecd.org/), Statista (https://www.statista.com/), Lista Firme (https://www.listafirme.ro/).

2.2 Collecting data using questionnaires

Design of a question and type of data generated

A more flexible tool for collecting data is the questionnaire. It allows researcher to customize the questions according to the purpose of the study and to generate the most convenient type of data.

Each question of the questionnaire is generating a variable.

A *bias* is the tendency of a process to evaluate a parameter under/over its value. When formulating questions, researcher must pay attention for avoiding *bias.*

Questions included in a questionnaire might be:

- **Closed questions** – respondent has options for choosing its answer(s). Formulating the question requires extra attention but working with generated data is more convenient.
- **Open questions** – responded has the freedom to formulate the answer but working with generated data is more difficult.

Design of the question is determining the type of data generated.

Example

Please appreciate the cleanness in your hotel room:

- **Unsatisfiable**
- **Satisfiable**
- **Good**
- **Very good**

is generating *qualitative data.*

On a scale from 1 (completely unsatisfied) to 10 (very satisfied) please evaluate the cleanness in your hotel room: 1 2 3 4 5 6 7 8 9 10

is generating *quantitative discrete data*.

On the following scale please mark with an X the cleanness in your hotel room:

completely unsatisfied highly satisfied

is generating *quantitative continuous data*.

How many employees does your company have? _____ is generating *quantitative discrete data.*

Which is your height? _____ is generating *quantitative continuous data.*

Traps when formulating a question

Researcher must be aware of the risk for a trap or haziness when formulating a question. Some of them are investigated below:

- **For unidirectional question, respondent has the tendency of choosing the favorable answer.**
  Is it difficult for graduates to have a bright future?
  - *Agree*
  - *Not agree*

  Question is evaluating only the difficulty of having a bright future, while the possibility of having it is ignored. Thus, it is a unidirectional question.

  Graduates will have a bright future
  - *Agree*
  - *Not agree*

  Question is evaluating only the existence of a bright future possibility of failure being ignored. Thus, it is a unidirectional question.

  It is recommendable to avoid unidirectional question. The two questions exemplified above can be reformulated as:

  Do you agree or not that it is difficult for graduates to have a bright future?

  Do you agree or not that those graduates will have a bright future?

- **When undecided answer is available, respondents have the tendency of choosing it.**
  Which is your opinion about the president?
  - *Favorable*
  - *Not favorable*
  - *I don't know*

  It is recommendable to eliminate the undecided answer *I don't know.*

- **For enumerating question, respondents have the tendency of choosing the first answer.**
  Please choose a color
  - *Blue*
  - *Red*

  It is recommendable to pay attention when choosing the order of enumeration.

- **When an anchor is present in the question, respondents have the tendency to remain close to it.**
  Knowing that US has a population of 316 mil, which is population of Canada?

  It is recommendable to avoid anchors, or if anchor might be somehow helpful it must be close to the real answer to avoid confusion.
  Question might be reformulated as:
  Which is the population of Canada?

- **It is recommendable to avoid complex questions which might confuse the respondent.**
  Do you consider that medical doctors and medical staff should benefit of special rights?

  This question is generating several possible categories: only medical doctors, only medical staff, both categories, no one of the two, thus creating a dilemma for the respondent: *which category to choose.*
  To avoid ambiguity, the question might be formulated as:

  Who do you consider should benefit of special rights?
  - *Medical doctors*
  - *Medical staff*
  - *Both categories*
  - *No one of the two*

- **Question should not contain unknown information to respondent which is playing a key role in formulating the answer.**
  Do you consider that people with Body Mass Index greater than or equal to 30 should avoid watching TV daily?
  - *Yes*
  - *No.*

  What does it mean Body Mass Index? If the concept is critical for choosing an answer, it must be explained.

  The question might be reformulated as:

  BMI (Body Mass index) is a convenient rule to categorize a person as underweight (under 18.5), normal weight (18.6 to 24.9), overweight (25 to 29.9), moderately obese (30 to 34.9),

severely obese (35 to 39.9) and very severely obese (above 40) by dividing its weight to the square of its height.

Do you consider that people with BMI greater that or equal to 30 should avoid watching TV daily?

- *Yes*
- *No.*

Auxiliary questions (variables)

Ideally, a sample must be a miniature of the population. Unfortunately, this is not always possible, especially due to a low answering rate. As consequence, some population classes are underrepresented and others overrepresented, impacting accuracy of the study.

Weighting the answers is a method to correct the misrepresentation. Auxiliary questions are introduced in the questionnaire. Usually, they refer to demographic factors (like gender, age, residency, marital status) with known distribution for the entire population. The auxiliary questions are generating the auxiliary variables fand their sample distribution is computed. For the auxiliary variables, the sample distribution is compared with known population distribution.

If the two distributions are comparable, then all population classes are well represented in the sample. Differences between the two distributions indicate misrepresentation of population classes in the sample. Weights calculated for the auxiliary variables are computed and used to correct the misrepresentation.

Example

A survey is analyzing income, color of the car and age. A questionnaire is applied for a sample of 10 persons.

Questions formulated in the questionnaire are

1. Which is your monthly income?
2. Which is the color of your car?
3. Which is your age category:
   - Young (less than or equal to 30)
   - Adult (31 to 60)
   - Senior (greater than or equal to 61)

and data collected are

Income	*Color of car*	*Age group*
2000	Blue	Y
2500	Red	Y
5000	Red	A
3000	Blue	S
3500	Green	S
5800	Blue	A
2900	Blue	Y
3200	Green	A
4200	Red	Y
6100	Red	S

Age category might be considered an auxiliary variable.

Its distribution for the entire population is considered known, being

Age category (population)	**%**
Y	30%
A	50%
S	20%

and its sample distribution is determined as being

Age category (sample)	**%**
Y	40%
A	30%
S	30%

Differences between the two distributions of age category are visible, indicating that misrepresentation of some population classes is present. More precise, young people *(40% in the sample vs 30% in population)* and seniors *(30% in the sample vs 20% in population)* are overrepresented while adults are underrepresented *(30% in the sample vs 50% in population).* Weighting the answers obtained from the sample is required. Importance of overrepresented classes should be reduced while for underrepresented classes it should be increased. Thus, weight is calculated by reporting % of population to the % of sample.

Weights calculated for each class are

Age category	**weights**
Y	$^{0.3}/_{0.4} = 0.75$
A	$^{0.5}/_{0.3} = 1.67$
S	$^{0.2}/_{0.3} = 0.67$

Using these calculated weights, the weighted values for income become

age category	weight	income	weighted income
Y	***0.75***	2000	2000**0.75* = 1500
Y	***0.75***	2500	2500**0.75* = 1875
A	***1.67***	5000	5000**1.67* = 8350
S	***0.67***	3000	3000**0.67* = 2010
S	***0.67***	3500	3500**0.67* = 2345
A	***1.67***	5800	5800**1.67* = 9686
Y	***0.75***	2900	2900**0.75* = 2175
A	***1.67***	3200	3200**1.67* = 5344
y	***0.75***	4200	4200**0.75* = 3150
S	***0.67***	6100	6100**0.67* = 4087

and the weighted value for color become

age category	color	weight
Y	Blue	0.75
Y	Red	0.75
A	Red	1.67
S	Blue	0.67
S	Green	0.67
A	Blue	1.67
Y	Blue	0.75
A	Green	1.67
Y	Red	0.75
S	Red	0.67
		10.02

generating the following distributions (original and weighted)

color	*original %*	*weighted %*
blue	40% = (1+1+1+1)/10	38.32% = (*0.75+0.67+1.67+0.75*)/10.02
red	40% = (1+1+1+1)/10	38.32% = (0.75+1.67+0.75+0.67)/10.02
green	20% = (1+1)/10	23.36% = (0.67+1.67)/10.02
	100%	100%

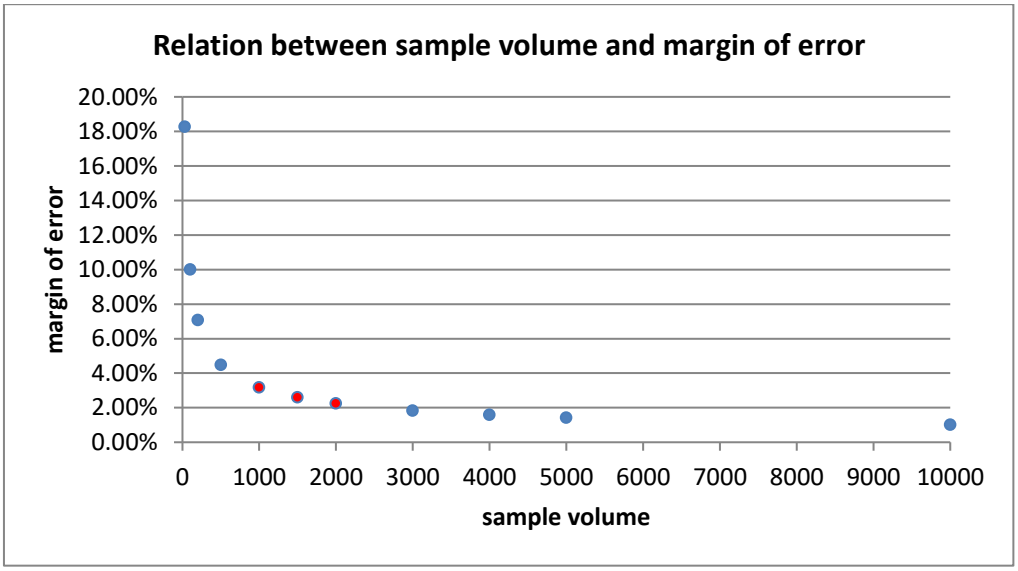
2.3 Sampling

2.3.1 Margin of error

For a given sample, the statistic is computed. We know that probably the computed sample statistics will not meet the real population parameter. Using another sample, it is possible to get a different statistic, which again will probably not meet the real population parameter. The

maximum difference which might occur between the real population parameter and the sample statistics is known as the margin of error.

Margin of error does not depend on the population volume (N), but only on sample volume (n). Margin of error is inverse related to sample volume (n) as it is visible in the following chart



Descending rate for margin of error is significant for small samples and almost irrelevant for very big samples. Big samples generate accurate results against a significant cost. Thus, a cost accuracy tradeoff is required. Gallup and Pew Research Center, probably the world survey leaders, are using samples with a volume ranging between 1000 and 2000.

Results provided by data analysis have a certain confidence level, the usual values being 90%, 95% or 99%. A margin of error of 3 obtained with a confidence level of 95% means that in 95 cases out of 100, the difference between real population parameter and computed sample statistic lies between +/- 3 and in 5 cases out of 100 the difference is exceeding the limit of +/- 3.

2.3.2 Sampling methods

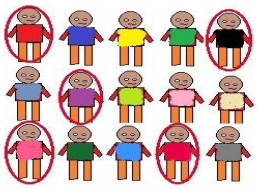
Sample is a miniature representation for the population. As consequence, elements selected for the sample should represent the entire population. There are two methods for sample selection.

1. **Probabilistic methods** - population elements to be included in the sample are randomly selected, probability of selecting each element being known. Probabilistic methods are complex, but they generate representative samples and accurate results.
2. **Non-probabilistic methods** – population elements to be included in the sample are subjectively choose by the researcher. There is a high chance for the sample not to accurate represent the population and *bias* might be significant.

Probabilistic methods

Simple Random Sampling

Each population element has the same probability of being selected in the sample and selection is performed randomly.



A graphic visualization for simple random sampling is available at the following link:

https://www.youtube.com/watch?v=yx5KZi5QArQ

Systematic Sampling

First element to be included in the sample is randomly selected from the population. A *step* is calculated (N/n) and the other sample elements are determined by applying the step to the first sample element and iterating the process.

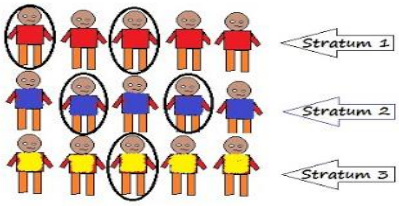
A graphic visualization for systematic sampling is available at the following link:

https://www.youtube.com/watch?v=QFoisfSZs8I

Stratified Sampling

Population is divided in groups called *strata.* Strata are heterogeneous (different) an all elements within stratum are homogenous (similar). Using random selection, elements from each stratum are selected for the sample.

Sample contains elements selected from <u>each</u> stratum.



A graphic visualization for stratified sampling sampling is available at the following link:

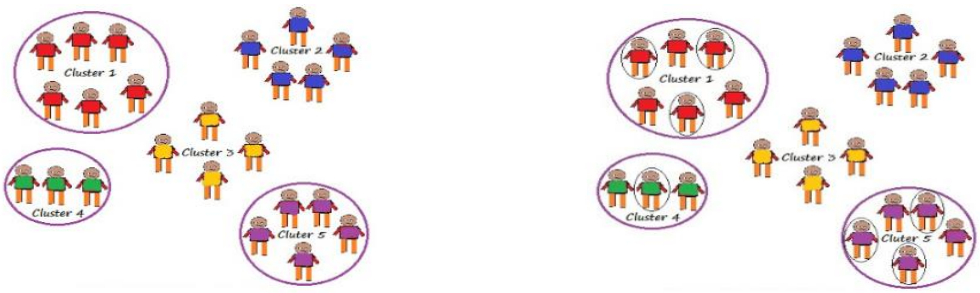
https://www.youtube.com/watch?v=sYRUYJYOpG0

Cluster Sampling

Population is divided in groups called *clusters.* Elements within each cluster are heterogenous (different), each cluster being a miniature of the population. Some of the clusters are randomly selected for generating the sample, two methods being possible:

- All elements of the selected clusters are included in the sample
- From the selected clusters, elements are randomly selected for being included in the sample

Sample contains elements selected from some of the clusters.



A graphic visualization for cluster sampling is available at the following link:

https://www.youtube.com/watch?v=QOxXy-I6ogs

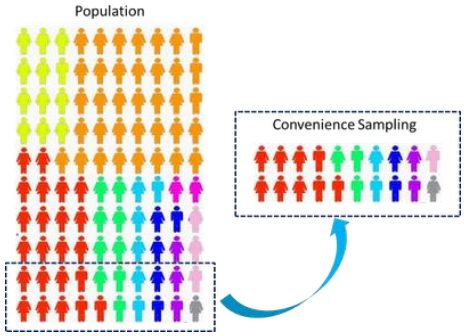
Non-probabilistic methods

Convenience sampling

Analyst is selecting the most accessible elements for being included in the sample. Is the leas complicated sampling method but bias is significant. A lot of aspects are out of control for researcher and credibility of studies employing this method is reduced.

Method is useful for testing a questionnaire or formulating a hypothesis.

For example, the analyst might ask all his contacts to fill in a questionnaire.

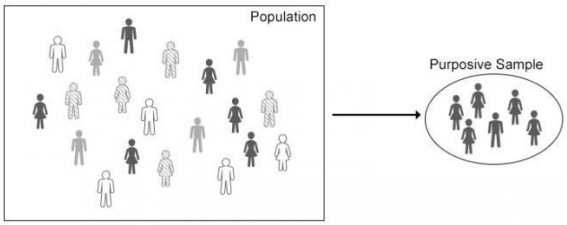


A graphic visualization for convenience sampling is available at the following link:

https://www.youtube.com/watch?v=aomNbRO5Zac

Purposive/judgement sampling

Analyst is using his own experience and knowledge to define a criterion for selecting elements to be included in the sample. Is one of the most efficient (time and cost) sampling methods, useful when access to population is limited. Method is sensitive to analyst judgement errors and might face a significant bias.



Purposive Sampling

A graphic visualization for purposive/judgement sampling is available at the following link:

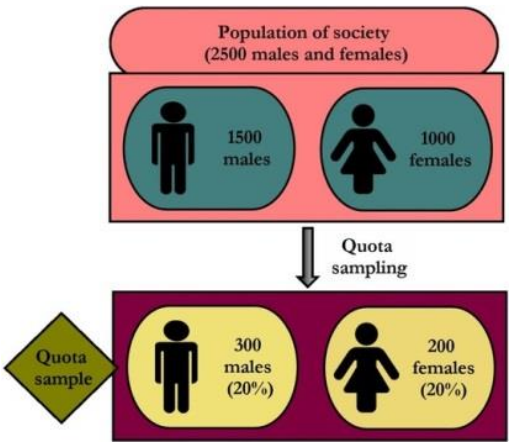
https://www.youtube.com/watch?v=CdK7N_kTzHI

Quota sampling

The analyst is dividing the population into groups using relevant criterion for the study like age, gender, income. Analyst is deciding the percentage of the population to be included in the sample and thus is determining how many elements from each group will be considered into the sample. Further analyst will choose in a subjective manner (not randomly) which elements from each group will be included in the sample.

Method is cost friendly and might be used when limited time is allocated for the study. As any other non-probabilistic method risk of bias is significant.

Stratified and quota sampling are similar, difference being determined by the way sample elements are chosen – random for stratified sampling and subjective for quota sampling.



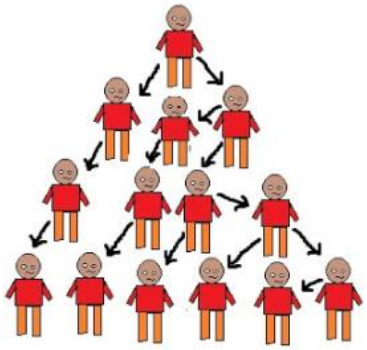
A graphic visualization for quota sampling is available at the following link:

https://www.youtube.com/watch?v=K8lcSHlB64w

Snowball sampling

Is a sampling method used when dealing with a sensitive subject (drugs consumption, chronic pain). Population cannot be identified, and sample elements are selected based on references provided by previous selected elements.

Is a highly efficient method (time and cost) but there is no control regarding sample representativity.



A graphic visualization for quota sampling is available at the following link:

https://www.youtube.com/watch?v=lq8dQel2ZRI

Sampling methodologies used by Gallup and Pew Research are available at the following links:

https://media.gallup.com/PDF/FAQ/HowArePolls.pdf

https://www.journalism.org/wp-content/uploads/sites/8/2020/03/PJ_2020.03.18_Coronavirus-News1_METHODOLOGY.pdf

3. Organization and presentation of data

The following questionnaire

Questionnaire

1. Please indicate the county of residence for your company.
2. Please indicate the no of employees in your company.
3. Please indicate the turnover of your company in the last year.

is applied on a sample of 50 entrepreneurs, collected data being

Questionnaire	*County*	*Employees*	*Turnover*	*Questionnaire*	*County*	*Employees*	*Turnover*
1	*CJ*	*4*	*916930*	*26*	*BH*	*3*	*97521*
2	*BN*	*4*	*97406*	*27*	*BN*	*8*	*768886*
3	*CJ*	*5*	*757557*	*28*	*AB*	*1*	*974347*
4	*BH*	*3*	*467018*	*29*	*BH*	*7*	*291942*
5	*BH*	*3*	*588042*	*30*	*AB*	*6*	*490498*
6	*BH*	*5*	*503741*	*31*	*CJ*	*8*	*560752*
7	*CJ*	*7*	*352155*	*32*	*AB*	*3*	*252879*
8	*AB*	*1*	*97437*	*33*	*MM*	*4*	*882307*
9	*BN*	*7*	*318087*	*34*	*AB*	*1*	*163690*
10	*BH*	*5*	*88920*	*35*	*MM*	*8*	*789122*
11	*CJ*	*4*	*975625*	*36*	*BN*	*7*	*499762*
12	*CJ*	*3*	*585501*	*37*	*CJ*	*0*	*960688*
13	*BN*	*1*	*375912*	*38*	*BN*	*1*	*959054*
14	*MM*	*6*	*940819*	*39*	*CJ*	*9*	*406511*
15	*AB*	*0*	*427737*	*40*	*AB*	*8*	*573812*
16	*CJ*	*1*	*69592*	*41*	*MM*	*2*	*870454*
17	*MM*	*1*	*703083*	*42*	*MM*	*7*	*741177*
18	*MM*	*9*	*795067*	*43*	*AB*	*2*	*388131*
19	*BN*	*3*	*269978*	*44*	*BN*	*7*	*767443*
20	*CJ*	*0*	*398641*	*45*	*BH*	*5*	*417183*
21	*CJ*	*5*	*124175*	*46*	*BH*	*8*	*702018*
22	*AB*	*3*	*990722*	*47*	*CJ*	*8*	*193391*
23	*MM*	*8*	*305280*	*48*	*CJ*	*8*	*108470*
24	*BH*	*3*	*760408*	*49*	*BH*	*9*	*600078*
25	*MM*	*3*	*508740*	*50*	*BH*	*6*	*751521*

Presenting collected data as a long que is not efficient, situation being even more visible when dealing with large samples.

Frequency tables and charts are efficient methods of presenting collected data.

Frequency tables might present data corresponding to a single variable (tabulation)

$$variabile: \begin{pmatrix} unique\ values\ of\ the\ variable \\ frequency \end{pmatrix}$$

or data corresponding to two variables (pivot tables)

Var 2 / Var 1	**Unique values of Var 2**	**Total**
Unique values of Var 1	frequency	marginal frequency
Total	marginal frequency	sample volume

Frequency tables and charts are easy generated using dedicated software like Excel, Statgraphics, R, Stata. All analysis exemplified in our lecture are based on Excel.

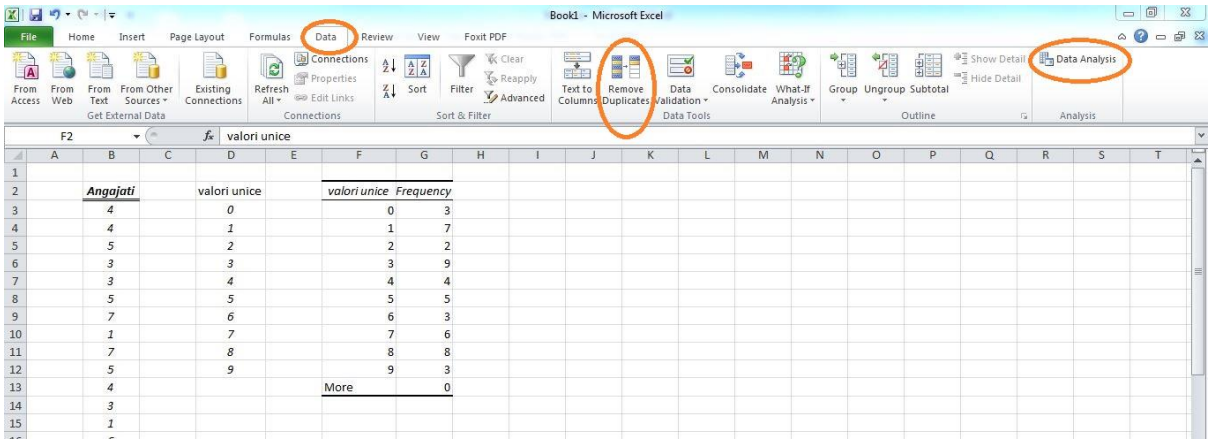
By default, **Data analysis** might not be configured in Excel. In this situation **Data analysis** option is not visible in the **Data** menu. To configure **Data analysis,** follow these steps:

- Access the **File** menu
- Select **Options**
- Select **Add-Ins**
- Identify and select **Analysis ToolPack** in the Name column and press **Go**
- Check **Analysis ToolPack** and press **OK.**

Tabulation of data

In our example we are dealing with 3 variables ***county*** – qualitative variable***, employees*** – quantitative discrete variable ***and turnover*** – quantitative continuous variable.

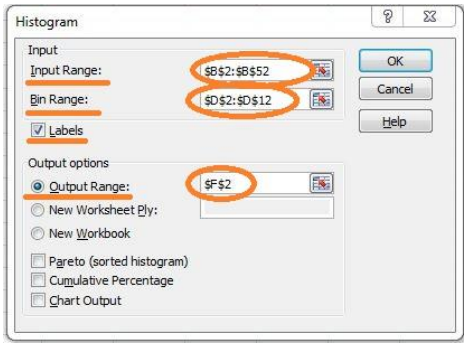
Tabulation of a *quantitative discrete variable* is a simple process in Excel. Let's presume that our quantitative discrete data *employee* is presented on column B in excel and labeled ***Angajati***



	B	D	F	G
2	Angajati	valori unice	valori unice	Frequency
3	4	0	0	3
4	4	1	1	7
5	5	2	2	2
6	3	3	3	9
7	3	4	4	4
8	5	5	5	5
9	7	6	6	3
10	1	7	7	6
11	7	8	8	8
12	5	9	9	3
13	4		More	0
14	3			
15	1			

Steps to follow for tabulation are

- **Generate the unique values** Copy Paste values on a new column (Column D) labeled **valori unice,** apply option **Data – Remove duplicates** to filter the unique values on column D and sort them ascending using the option **Home – Sort & Filter – Sort smallest to largest.**
- For **Tabulation** access **Data – Data analysis** and by selecting **Histogram** the following window becomes active



- Select **the analyzed data** (the 50 values to be analyzed – in our example they are on Column B) in the **Input Range** field.
- Select **the unique value** (in our example they are on Column D) in the **Bin Range** field.
- If data were selected together with their labels, then check the **Labels** field.
- Chose the output option for the frequency table from the 3 possible options:
  - Same worksheet starting with a certain cell (**output range**)
  - New worksheet
  - New workbook

  We choose to generate the table in the same worksheet starting with cell F2, thus we check **Output Range** and select F2

- Press **OK** and the frequency table is generated.

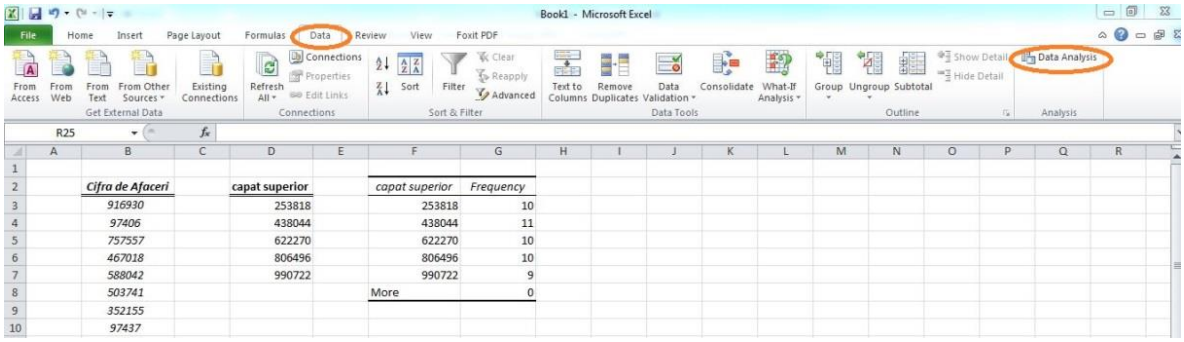
By default, the absolute frequencies (no of occurrence) are computed. Relative frequencies (percentage of occurrence) are easily computed by dividing each absolute frequency to the total.

The frequency table for the employees of the 50 analyzed companies is

employee	*abs freq*	*rel freq*
0	3	6%
1	7	14%
2	2	4%
3	9	18%
4	4	8%
5	5	10%
6	3	6%
7	6	12%
8	8	16%
9	3	6%

7 companies out of the 50 analyzed, meaning 14% of the total have only 1 employee.

Tabulation of a *quantitative continuous variable* is also a simple process in Excel. Let's presume that our quantitative continuous data *turnover* is presented on column B in excel and labeled ***Cifra de Afaceri.***



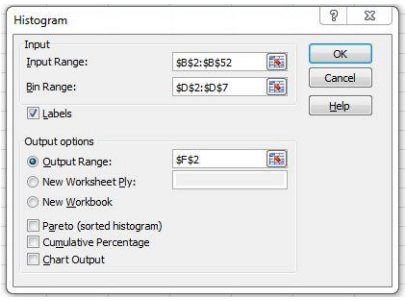
	Cifra de Afaceri	capat superior	capat superior	Frequency
3	916930	253818	253818	10
4	97406	438044	438044	11
5	757557	622270	622270	10
6	467018	806496	806496	10
7	588042	990722	990722	9
8	503741		More	0
9	352155			
10	97437			

Steps to follow for tabulation are

- Continuous data are very diverse, thus there will be several unique values and presenting them separate is not an efficient option. Grouping data in intervals is the efficient way to present them. Usually intervals are equidistant (have the same length). **Superior margin of each interval must be computed.** The analyst is deciding the number of intervals and calculating the interval length as

$$l = \frac{maximum\ value - minimum\ value}{no\ of\ intervals}$$

  Intervals are computed in an iterative process and superior margin of each interval is registered on **Column D** and labeled ***capat superior.***
- For **Tabulation** access **Data – Data analysis** and by selecting **Histogram** the following window becomes active



- Select **the analyzed data** (the 50 values to be analyzed – in our example they are on Column B) in the **Input Range** field.
- Select **superior margins of intervals** (in our example they are on Column D) in the **Bin Range** field.
- If data were selected together with their labels, then check the **Labels** field.
- Chose the output option for the frequency table from the 3 possible options:
  - Same worksheet starting with a certain cell (**output range**)
  - New worksheet
  - New workbook

  We choose to generate the table in the same worksheet starting with cell F2, thus we check **Output Range** and select F2

- Press **OK** and the frequency table is generated.

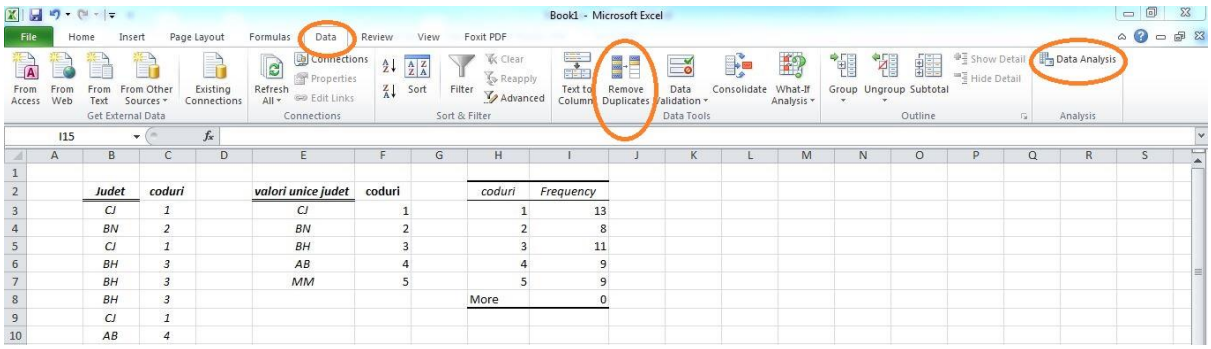
By default, the absolute frequencies (no of occurrence) are computed. Relative frequencies (percentage of occurrence) are easily computed by dividing each absolute frequency to the total.

The frequency table for the turnover of the 50 analyzed companies is

turnover	*abs freq*	*rel freq*
[69592 - 253818]	10	20%
(253818 - 438044]	11	22%
(438044 - 622270]	10	20%
(622270 - 806496]	10	20%
(806496 - 990722]	9	18%

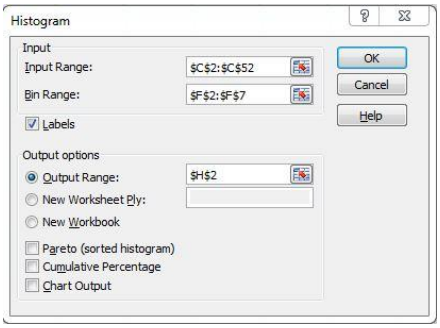
11 companies out of the 50 analyzed, representing 22% of the total have a turnover between 253818 and 438044 lei.

Tabulation of a *qualitative variable* is a bit complicated process in Excel, because Excel is counting only numbers and not symbols. Let's presume that our qualitative data *county* is presented on column B in excel and labeled ***judet.***



Steps to follow for tabulation are

- **Generate the unique values** Copy Paste values on a new column (Column D) labeled **valori unice** and apply option **Data – Remove duplicates** to filter the unique values on column D.
- **Generate the codes** Excel does not count symbols, thus a numeric code must be attached to each qualitative data. For each unique value a numeric code must be defined. In our example these codes are introduced on Column F. Using for example the VLOOKUP function in Column C, the corresponding numeric code is attached to each sample value. Further, only codes will be employed.
- For **Tabulation** access **Data – Data analysis** and by selecting **Histogram** the following window becomes active



- Select **the codes corresponding to the analyzed data** (the 50 corresponding codes are on Column C in our example) in the **Input Range** field.
- Select **the unique codes** (in our example they are on Column F) in the **Bin Range** field.
- If codes were selected together with their labels, then check the **Labels** field.
- Chose the output option for the frequency table from the 3 possible options:
  - Same worksheet starting with a certain cell (**output range**)
  - New worksheet
  - New workbook

  We choose to generate the table in the same worksheet starting with cell H2, thus we check **Output Range** and select H2

- Press **OK** and the frequency table is generated.

By default, the absolute frequencies (no of occurrence) are computed. Relative frequencies (percentage of occurrence) are easily computed by dividing each absolute frequency to the total.

The frequency table for the county of residence of the 50 analyzed companies is

county	*abs freq*	*rel freq*
AB	9	18%
BH	11	22%
BN	8	16%
CJ	13	26%
MM	9	18%

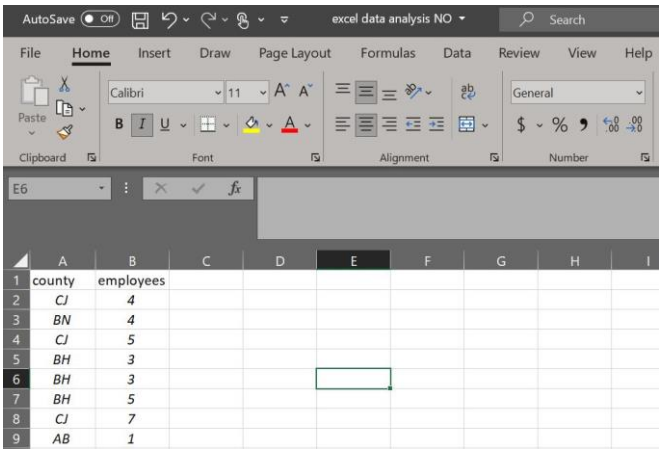
11 companies out of the 50 analyzed, representing 22% of the total have their headquarter in Bihor County.

Pivot tables

Pivot tables are analyzing the joint frequencies for two variables, opening the path for new analysis (correlation, regression). Pivot tables are easily computed in Excel by using the **Insert – Pivot Table** option. Latest versions of Excel offer the facility of simultaneous computing both Pivot Table and Pivot Chart (**Insert – Chart – Pivot Chart – Pivot Chart & Pivot Table** option)

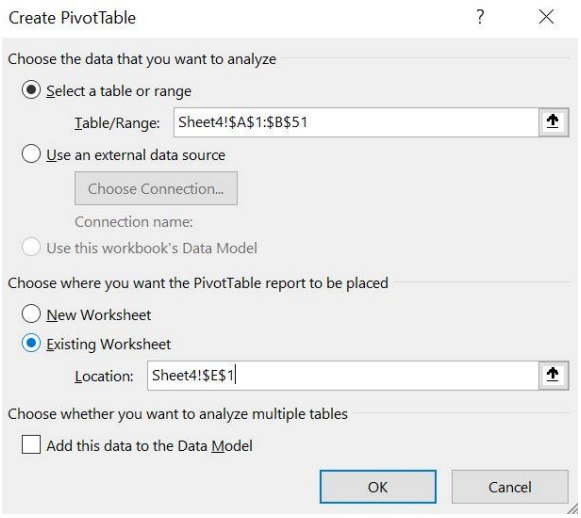
Extra attention must be provided when dealing with quantitative continuous data which usually have a wide diversity. To avoid counting each individual value separate, instead of working with intervals, we must code each interval and attach the corresponding interval code to each sample value. Further codes will be used instead of sample values.

Steps for generating the pivot table will be explained using variables county and employee. Values of the two variables are introduced on two separate columns, Columns A for county and B for employee.



	A	B	C	D	E	F	G	H	I
1	county	employees							
2	CJ	4							
3	BN	4							
4	CJ	5							
5	BH	3							
6	BH	3							
7	BH	5							
8	CJ	7							
9	AB	1							

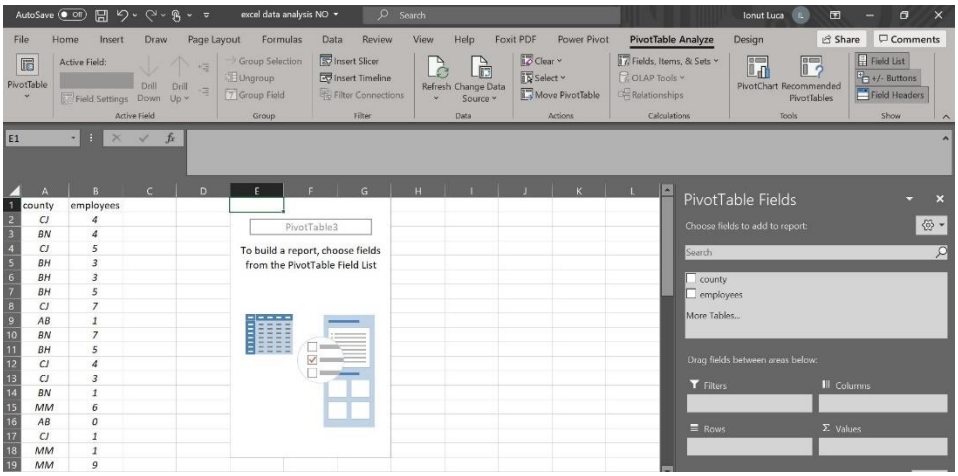
By selecting Pivot Table in menu Insert, the following window pops-up.



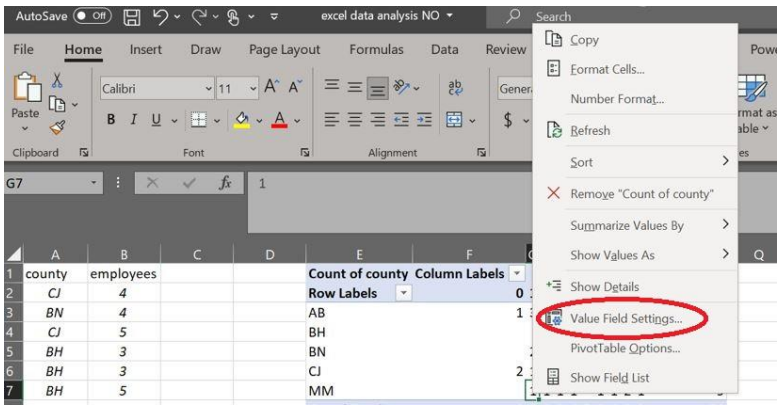
Field **Table/Range** must be filled in with data to be analyzed (values for county and employee, including the labels).

Location for generating the pivot table must be indicated. We chose to generate it in the same sheet, starting with cell E1.

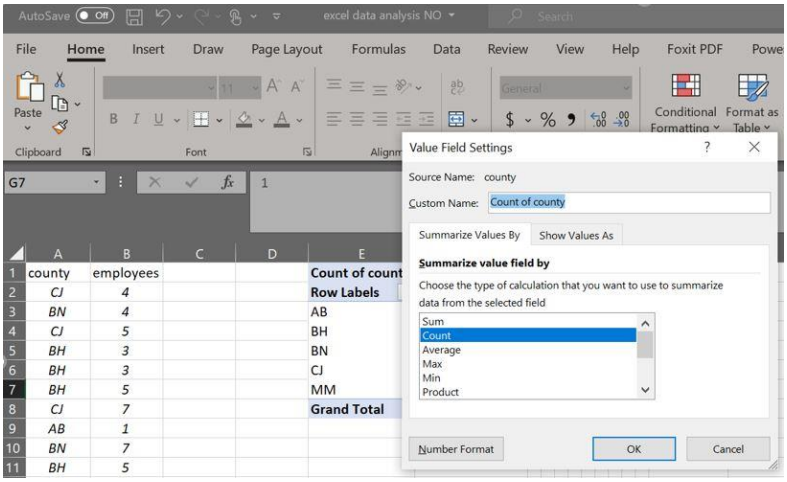
Press OK and the following window pops-up.



To generate the pivot table variables must be allocated on rows and columns. The analyst is deciding which variable is allocated to row and which to column, according to design of research. *In our case we allocate county to row and employee to column*. To determine the frequencies, one of the two variables (analyst is deciding which one) must be allocated to the field values. *In our case we allocate employee to the field values.* From all possible settings of field Value, count must be selected. For it, right click on the pivot table, select Value Field Settings



and then choose Count



The pivot table for employee vs county is

Count of county	Column Labels										
Row Labels	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**Grand Total**
AB	1	3	1	2			1		1		9
BH				4		3	1	1	1	1	11
BN		2		1	1			3	1		8
CJ	2	1		1	2	2		1	3	1	13
MM		1	1	1	1		1	1	2	1	9
Grand Total	**3**	**7**	**2**	**9**	**4**	**5**	**3**	**6**	**8**	**3**	**50**

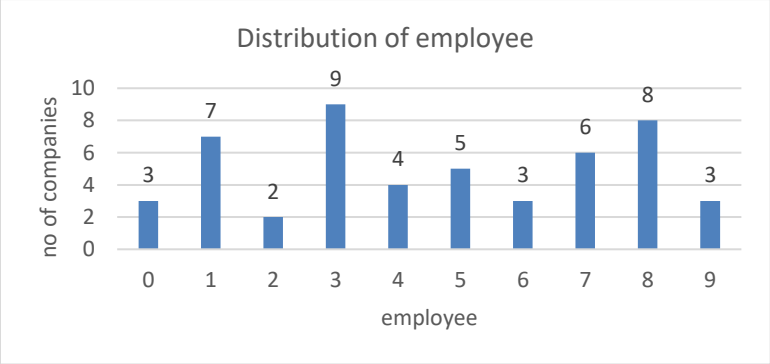
4 companies out of the 5 analyzed have 3 employee and headquarter in Bihor County.

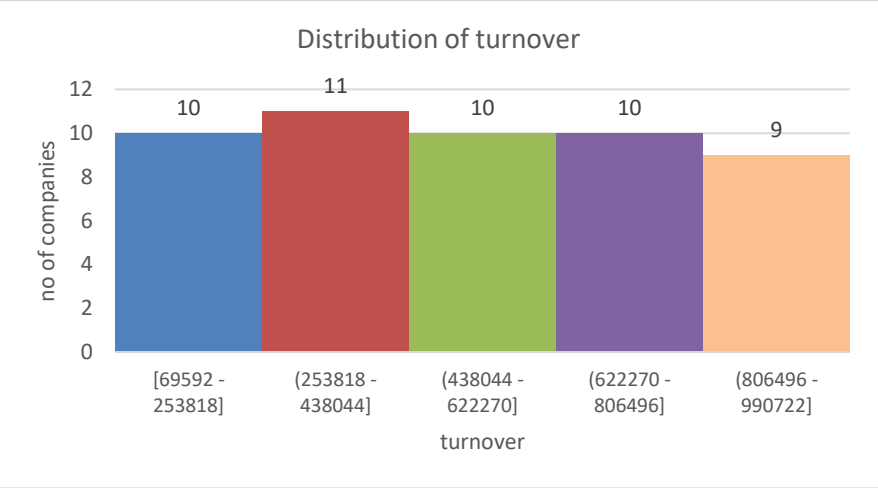
Charts

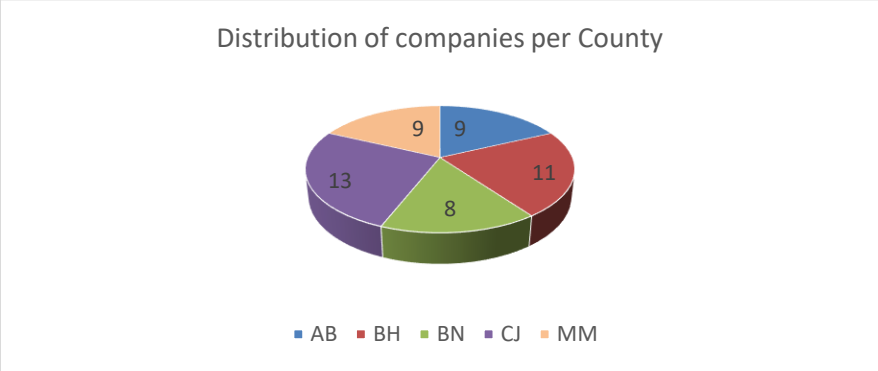
Frequently used charts for presenting data are histogram, pie and pivot chart.

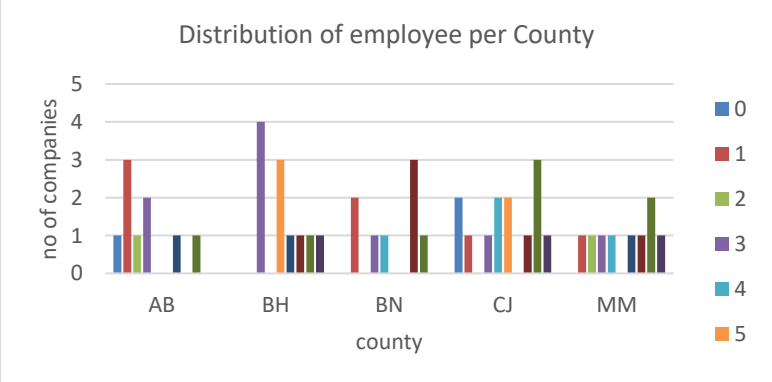
For quantitative continuous data we recommend histogram (emphasizing continuity and trend), while for qualitative data we recommend pie chart (emphasizing weight in the total). For quantitative discrete data chart should relate to what we intend to emphasize. If trend is in our focus, then we should choose histogram and if weight is in our focus, then we should choose pie chart.

Examples of charts for employee, turnover, county, and county vs employee are


Distribution of employee
no of companies
employee
0 1 2 3 4 5 6 7 8 9
3 7 2 9 4 5 3 6 8 3


Distribution of turnover
no of companies
turnover
[69592 - 253818]
(253818 - 438044]
(438044 - 622270]
(622270 - 806496]
(806496 - 990722]
10 11 10 10 9


Distribution of companies per County
9 11 8 13 9
AB BH BN CJ MM


Distribution of employee per County
no of companies
county
AB BH BN CJ MM
0 1 2 3 4 5

4. Data analysis

4.1 Indexes

Analyzed data might refer to time or space distribution of a phenomenon. Measuring changes from one class to another might be useful. Let us consider the following distribution of data

$$Y: \begin{pmatrix} y_1 & y_2 & y_3 & \dots & y_{n-2} & y_{n-1} & y_n \\ f_1 & f_2 & f_3 & & f_{n-2} & f_{n-1} & f_n \end{pmatrix}$$

Changes from one class to another is measured using indexes.

When reporting to previous class indexes are called **chain-based indexes** and when reporting to a fixed class (usually the first one) indexes are called **fix-based indexes**.

Main indexes are:

- **Absolute difference –** is analyzing how much (quantity) the analyzed phenomenon is changing compared to the reference class. Absolute difference is calculated as $\Delta = \boldsymbol{f_i} - \boldsymbol{f_1}$, $\forall \boldsymbol{i} = \overline{\boldsymbol{1,n}}$, for fix-based and as $\Delta = \boldsymbol{f_i} - \boldsymbol{f_{i-1}}$, $\forall \boldsymbol{i} = \overline{\boldsymbol{1,n}}$ for chain-based.
  Negative absolute difference is indicating a decrease and positive absolute difference is indicating an increase.
- **Relative difference** – is analyzing how much (percentage) the analyzed phenomenon is changing compared to the reference class. Relative difference is calculated as $\boldsymbol{R} = \left(\frac{f_i}{f_1} - \boldsymbol{1}\right) \times \boldsymbol{100}$, $\forall \boldsymbol{i} = \overline{\boldsymbol{1,n}}$ for fix-based and as $\boldsymbol{R} = \left(\frac{f_i}{f_{i-1}} - \boldsymbol{1}\right) \times \boldsymbol{100}$, $\forall \boldsymbol{i} = \overline{\boldsymbol{1,n}}$ for chain-based.
  Negative relative difference is indicating a decrease and positive relative difference is indicating an increase.
- **Specific weight –** is indicating weight of a class in the total, being calculated as $\boldsymbol{g_i} = \frac{f_i}{\boldsymbol{total}} \times \boldsymbol{100}$, $\boldsymbol{i} = \overline{\boldsymbol{1,n}}$.

To exemplify the indexes, let's analyze monthly sales of a company in 6 locations.

$$mth\ sales: \begin{pmatrix} Jan & Feb & Mar & Apr & May & Jun & Jul & Aug & Sep & Oct & Nov & Dec \\ 6500 & 7000 & 7300 & 8000 & 8200 & 9000 & 10000 & 8000 & 7800 & 7500 & 7200 & 6800 \end{pmatrix}$$

$$sales/location: \begin{pmatrix} Bihor & Brasov & Cluj & Mures & Sibiu & Alba \\ 15000 & 16000 & 18000 & 16500 & 15000 & 12800 \end{pmatrix}$$

Indexes for monthly sales are

mth	sales	absolute difference		relative difference		specific weight
		fix-based	chain-based	fix-based	chain-based	
Jan	6500	0	x	0.00%	X	7%
Feb	7000	500	500	7.69%	7.69%	8%
Mar	7300	800	300	12.31%	4.29%	8%
Apr	8000	1500	700	23.08%	9.59%	9%

May	8200	1700	200	26.15%	2.50%	9%
Jun	9000	2500	800	38.46%	9.76%	10%
Jul	10000	3500	1000	53.85%	11.11%	11%
Aug	8000	1500	-2000	23.08%	-20.00%	9%
Sep	7800	1300	-200	20.00%	-2.50%	8%
Oct	7500	1000	-300	15.38%	-3.85%	8%
Nov	7200	700	-300	10.77%	-4.00%	8%
Dec	6800	300	-400	4.62%	-5.56%	7%

Sales of March have increased with 800 € compared to January, representing an increase with 12.31%. Sales of March have increased with 300 € compared to February, representing an increase with 4.29%.
Sales of August have increased with 1500 € compared to January, representing an increase with 23.08%.
Sales of August have decreased with 2000 € compared to July, representing a decrease with 20%.
10% of the total yearly sales have been realized in June.

Indexes for sales per location are

location	sales	absolute difference		relative difference		specific weight
		fix-based	chain-based	fix-based	chain-based	
Bihor	15000	0	x	0.00%	X	16%
Brasov	16000	1000	1000	6.67%	6.67%	17%
Cluj	18000	3000	2000	20.00%	12.50%	19%
Mures	16500	1500	-1500	10.00%	-8.33%	18%
Sibiu	15000	0	-1500	0.00%	-9.09%	16%
Alba	12800	-2200	-2200	-14.67%	-14.67%	14%

Sales in Cluj are higher than in Bihor with 3000 € (20%).
Sales in Mures are lower than in Cluj with 8.33% (1500 €).
Sales in Sibiu and Bihor are similar.
Cluj has realized 19% of total sales performed in the 6 analyzed locations.

4.2 Statistics

Statistics explain central tendency of data, their spread around central tendency and their distribution, creating a helicopter view for sample data.

Average is probably the most used and well-known statistics. Average indicates the hypothetic value of data if all the influence factors would act the same on data.

Example: Grades of 5 students at math exam are: 9, 10, 5, 8, 7. The average grade at math exam is 7.8. If all the students would (1) attend the lectures and seminars, (2) allocate the same number of

hours for study, (3) have the same passion for math, then each one of the five students would have grade of 7.8 at math exam. Considering that these influence factors don't act the same for the five analyzed students, their grades at math exam are not all equal to 7.8.

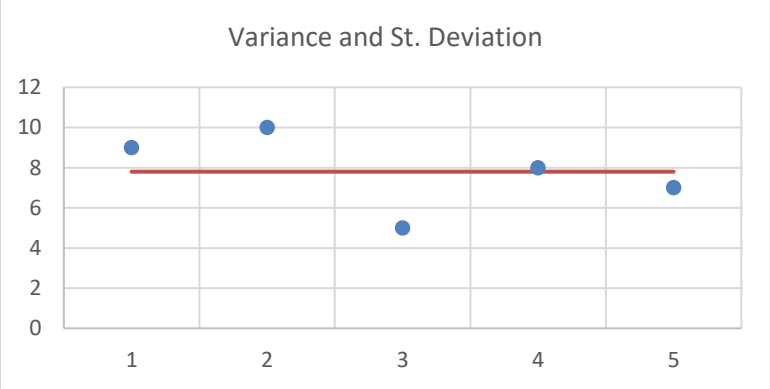
Average is easily computed with Excel, using the formula

=AVERAGE (number 1, number 2, …)

Average might be calculated only for quantitative data.

Variance, Standard deviation, and Coefficient of Variation

As it was visible in the example for average, data fluctuate around the *average*. Magnitude of this fluctuation is an important statistic, necessary and useful to be computed.



Variance is measuring the magnitude of data fluctuation around average. When calculating this magnitude, each individual fluctuation is squared up to avoid compensation between negative and positive values. As consequence of square up, the measuring unit is altered, making it difficult to understand and interpret variance.
To correct this inconvenient, variance is square rooted, generating **standard deviation**.
Both variance and standard deviation measure the same fluctuation, one having an inconvenient measuring unit (variance) and the other one (st. deviation) having a convenient measuring unit and being more user friendly.

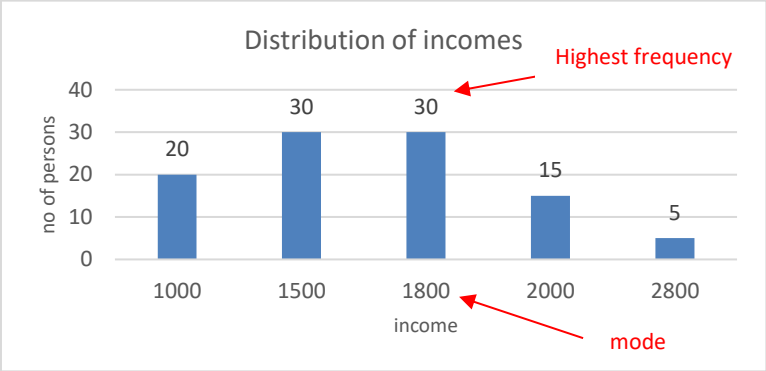
Analyzing and isolated number is impossible to say if that number is big or small. It's necessary to compare it with another number to decide if it's small or big. This principle is applied also for variance/st. deviation.
Coefficient of variation is comparing st. deviation with average and helps indicating if fluctuation of data around average is small or not.

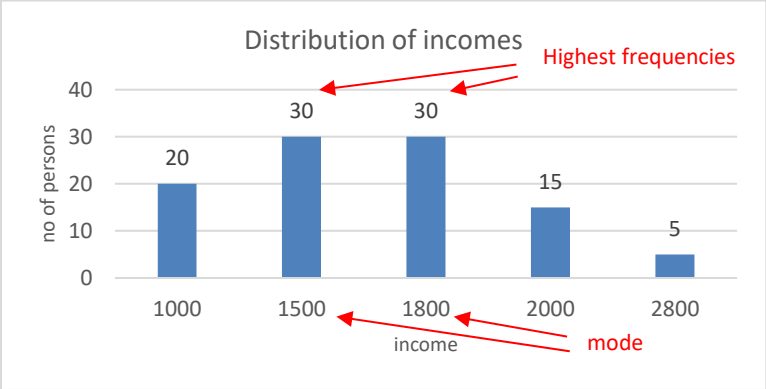
Variance, st. deviation and coef. of var are calculated in Excel using the formulas

=VAR (number1, number 2, …)	for variance
=STDEV (number1, number 2, …)	for st. deviation
= STDEV / AVERAGE	for coef. of var

Variance, st. deviation and coef. of var might be calculated only for quantitative data.

Mode is indicating the most frequent value. When analyzing data, it's possible to have a single value with the highest frequency (this value is the mode) or is possible to have several values with the highest frequency, each one being mode.





Mode might be determined directly from the distribution, by identifying that class (those classes) corresponding to the highest frequency (or frequencies), or in Excel by using the formula

=MODE (number1, number2, …)

In case of data with several mode values, the Excel formula will return #N/A, meaning that computing was not possible. Mode value must be determined from the distribution in case of discrete data (qualitative or quantitative) or must be approximated in case of continuous data, using the formula

$$Mo = x_{k-1} + \frac{\Delta_1}{\Delta_1 + \Delta_2} l$$

where x_{k-1} is the smallest value of the mode interval (that interval from the distribution having the highest frequency).

$\Delta_1 = f_k - f_{k-1}$ (frequency of the mode interval – frequency of the previous interval)

$\Delta_2 = f_k - f_{k+1}$ (frequency of the mode interval – frequency of the next interval)
$l = x_k - x_{k-1}$ (length of the mode interval)

Mode might be calculated both for qualitative and quantitative data. Unfortunately, Excel can't count qualitative data. To handle this shortcoming of Excel, we might use codes. Thus, qualitative data are transformed in quantitative data, accepted as input for *MODE* formula in Excel.

Quartile are dividing data into 4 equal classes.
Q_1 or **first quartile** is delimitating the smallest 25% of the values.
Q_2 or **second quartile** or **median** is dividing values in 2 equal parts, half being smaller than median and half greater than median.
Q_3 or **third quartile** is delimitating the highest 25% of the values.

Quartiles are easily calculated in Excel using the following formulas

=QUARTILE (array,1) *for Q_1*
=QUARTILE (array,2) *for Q_2*
=QUARTILE (array,3) *for Q_3*

Quartile might be calculated only for quantitative data.

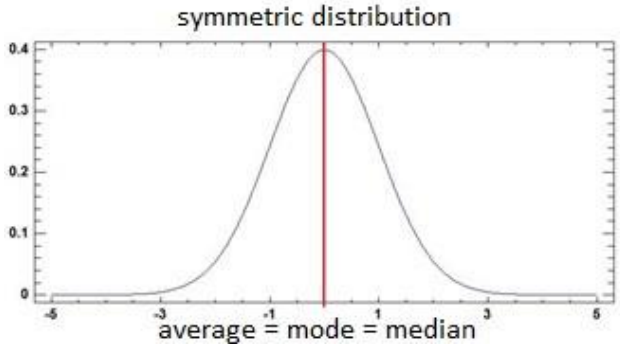
A brief inventory of the presented statistics, their notation, type of data for which they might be computed, and the computational formula implemented in Excel are presented in the following table:

Statistics	Notation	Type of data		Excel
		Qualitative data	Quantitative data	
Average	$\bar{x}$	✗	✓	*=AVERAGE (no1, no2, …)*
Variance	s^2	✗	✓	*=VAR (no1, no2, …)*
Standard deviation	s	✗	✓	*=STDEV (no1, no2, …)*
Coefficient of Variation	V_x	✗	✓	$= \frac{STDEV\ (no1, no2, \dots)}{AVERAGE\ (no1, no2, \dots)}$
Mode	Mo	✓	✓	*=MODE (no1, no2, …)*
First Quartile	Q_1	✗	✓	*=QUARTILE (array,1)*
Second Quartile (Median)	$Q_2; Me$	✗	✓	*=QUARTILE (array,2)*
Third Quartile	Q_3	✗	✓	*=QUARTILE (array,3)*

A few remarks regarding statistics are important and useful:

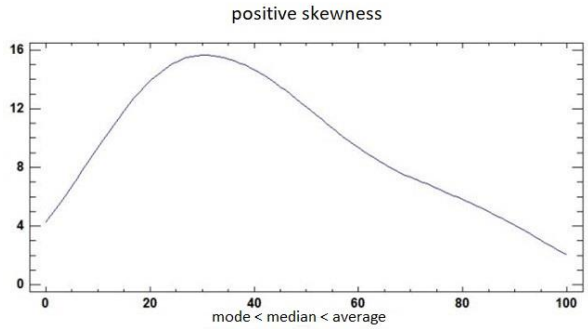
1. Average, Mode and Median are measures of central tendency of data.
2. Quartiles indicate the structure of data.
3. Variance, st. deviation, and coef. of var measure spread of data around central tendency.

4. In an ideal situation average, mode, and median are equal, indicating a symmetric distribution of data.

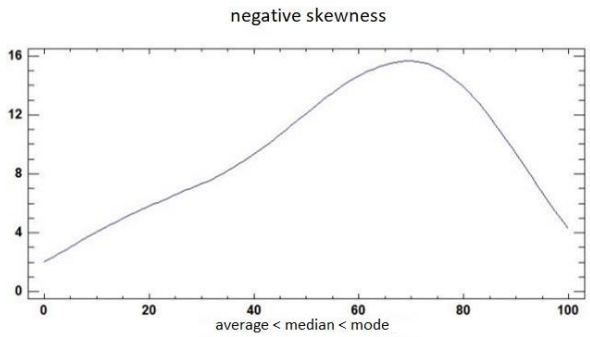


If average, mode, and median are not equal, then data are skewed, meaning that one tail of the distribution is longer than the other, being possible two scenarios:

- Positive skewness – distribution has a longer right tail, indicating that there are some "extremely high" values, generating an increase of the average.



- Negative skewness – distribution has a longer left tail, indicating that there are some "extremely small" values, generating a decrease of the average.



Using statistics, we will analyze the sample data obtained from the 50 companies.

The variable county contains qualitative data, thus the only statistics which might be calculated is *mode.*

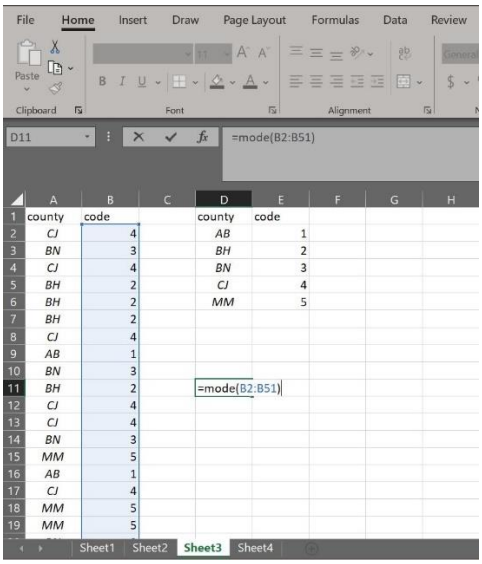
Using distribution of the 50 companies

county	*abs freq*	*rel freq*
AB	9	18%
BH	11	22%
BN	8	16%
CJ	13	26%
MM	9	18%

we identify the highest frequency as being 13, and thus the corresponding county CJ is the mode.
It means that most of the 50 analyzed companies have their residency in Cluj County.

To compute mode using the Excel formula, we must allocate codes for converting qualitative data into quantitative. The unique qualitative values of the variable are identified. Each unique qualitative value is receiving a numeric code. Analyst has the freedom to decide on the numeric code. Corresponding codes are allocated to each qualitative value of the variable – for efficiency VLOOKUP function might be employed. *MODE* function is employed for the codes, determining in cell D11 the mode value.

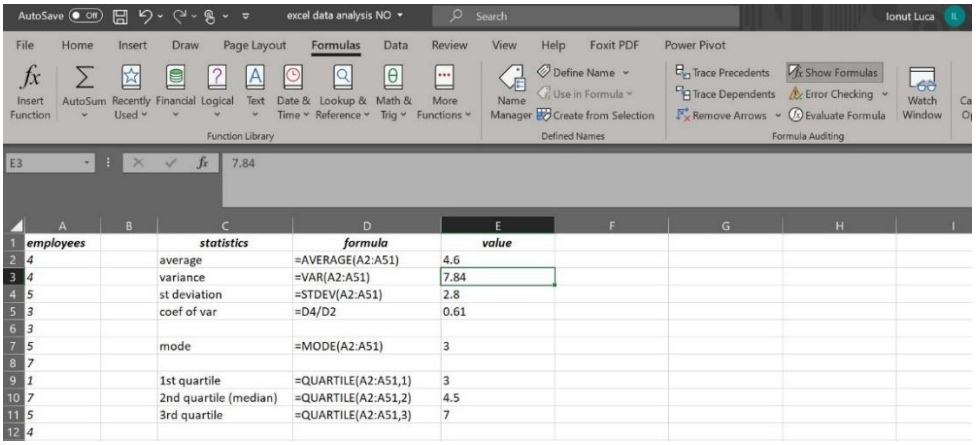
Synthesis of these steps is presented in the below diagram



MODE function is returning value 4, which is the code for Cluj County.
As consequence, the most of the 50 analyzed companies have their residency in Cluj County.

The variable employees and turnover contain quantitative data, thus all statistics might be computed.

Formulas applied for employees are displayed in the following diagram



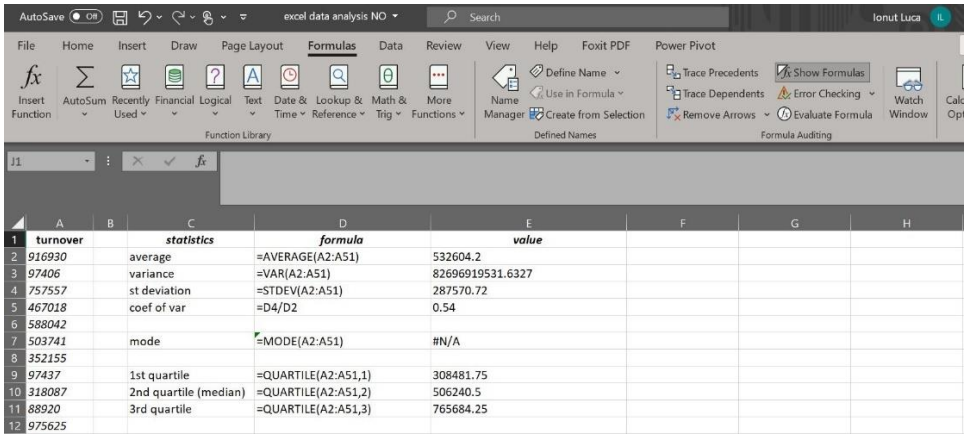
	A	B	C	D	E	F	G	H	I
1	*employees*		*statistics*	*formula*	*value*				
2	4		average	=AVERAGE(A2:A51)	4.6				
3	4		variance	=VAR(A2:A51)	7.84				
4	5		st deviation	=STDEV(A2:A51)	2.8				
5	3		coef of var	=D4/D2	0.61				
6	3								
7	5		mode	=MODE(A2:A51)	3				
8	7								
9	1		1st quartile	=QUARTILE(A2:A51,1)	3				
10	7		2nd quartile (median)	=QUARTILE(A2:A51,2)	4.5				
11	5		3rd quartile	=QUARTILE(A2:A51,3)	7				
12	4								

Meanings of the results are

statistics	***value***	***meaning***
average	4.6	The 50 companies have in average 4.6 employees.
variance	7.84	Measuring unit is squared persons, thus is not interpreted.
st. deviation	2.80	No of employees from the 50 companies, fluctuate from the average of 4.6 employees in average with 2.8 persons.
coef of var	61%	There is an important fluctuation for no of employees in the 50 companies
mode	3	The most of the 50 companies have 3 employees.
first quartile	3	¼ of the 50 companies have between 0 and 3 employees.
second quartile (median)	4.5	½ of the 50 companies have between 0 and 4.5 employees, while the other ½ of the 50 companies have between 4.5 and 9 employees.
third quartile	7	¼ of the 50 companies have between 7 and 9 employees.

The overall picture of the sample is: *The most of the 50 companies have a small number of employees, but there a companies with several employees, contributing to the increase of sample average. A little bit more that 50% of the companies are below average, but there exists a significant variability between companies.*

Formulas applied for turnover are displayed in the following diagram



	A	B	C	D	E	F	G	H
1	*turnover*		*statistics*	*formula*	*value*			
2	916930		average	=AVERAGE(A2:A51)	532604.2			
3	97406		variance	=VAR(A2:A51)	82696919531.6327			
4	757557		st deviation	=STDEV(A2:A51)	287570.72			
5	467018		coef of var	=D4/D2	0.54			
6	588042							
7	503741		mode	=MODE(A2:A51)	#N/A			
8	352155							
9	97437		1st quartile	=QUARTILE(A2:A51,1)	308481.75			
10	318087		2nd quartile (median)	=QUARTILE(A2:A51,2)	506240.5			
11	88920		3rd quartile	=QUARTILE(A2:A51,3)	765684.25			
12	975625							

Result presented for mode is #N/A, indicating that mode was not computed. A deeper analysis of data reveals that each data has frequency 1. We are dealing with quantitative continuous data, thus mode must be approximated. Let's recall the five intervals distribution previously computed

turnover	*abs freq*	*rel freq*
[69592 - 253818]	10	20%
(253818 - 438044]	11	22%
(438044 - 622270]	10	20%
(622270 - 806496]	10	20%
(806496 - 990722]	9	18%

Second interval has the highest frequency thus it is the mode interval (interval containing mode). Applying approximation formula, mode is approximated by

$$Mo = x_{k-1} + \frac{\Delta_1}{\Delta_1 + \Delta_2} l = 253818 + \frac{(11-10)}{(11-10)+(11-10)} * 184226 = 345931$$

Meanings of computed results for statistics are

statistics	***value***	***meaning***
average	532604.20	Average turnover of the 50 companies is 532604.20€
variance	82696919531.6	Measuring unit is €^{2}, thus is not interpreted.
st. deviation	287570.72	Turnover of the 50 companies is fluctuating around 532604.2€ in average with 287570.72€.
coef of var	54%	There is an important fluctuation for the turnover of the 50 companies
mode	345931	Majority of the 50 companies have a turnover of approximately 345931€
first quartile	308481.75	¼ of the 50 companies has a turnover smaller that 308481.75€. This is bounding companies with weak performances.
second quartile (median)	506240.50	½ of the 50 companies has a turnover below 506240.50€, while the other ½ of the 50 companies has a turnover above 506240.50€.
third quartile	765684.25	¼ of the 50 companies have a turnover above 765684.25€. This is bounding the highly performant companies.

The overall picture of the sample is: *Majority of the 50 companies have a small turnover, but there a companies with high turnover, contributing to the increase of sample average. More than 50% of the companies have a turnover below average, variability between companies being significant.*

4.3 Correlation of data

In case of two quantitative variables, there exists a simple technique exploring whether there exists an association between them. In our example we might analyze the effect of turnover on the

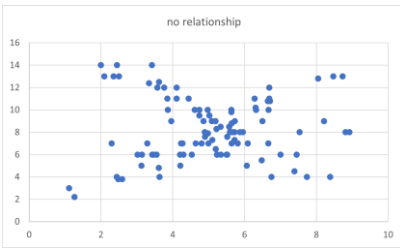
number of employees a company has. Thus, one variable is the explanatory one, while the other is the response variable.

Data corresponding to our sample of 50 companies are

	explanatory	response
1	turnover	*employees*
2	*916930*	*4*
3	*97406*	*4*
4	*757557*	*5*
5	*467018*	*3*
6	*588042*	*3*
7	*503741*	*5*
8	*352155*	*7*
9	*97437*	*1*
10	*318087*	*7*
11	*88920*	*5*
12	*975625*	*4*

The first step in exploring the relationship between turnover and employees is to create a scatterplot chart. This type of chart helps understanding the overall pattern (direction, shape and strength) and deviations from the pattern (outliers).

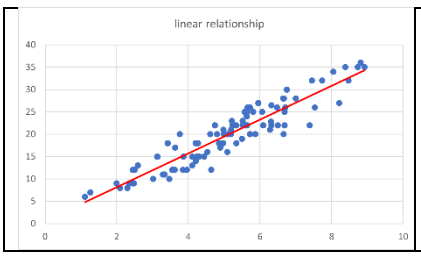
Possible **directions** of data are

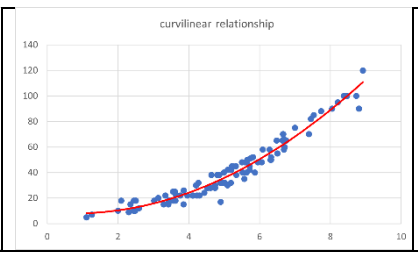


Positive relationship indicates that an increase of the explanatory variable is generating an increase of response variable.

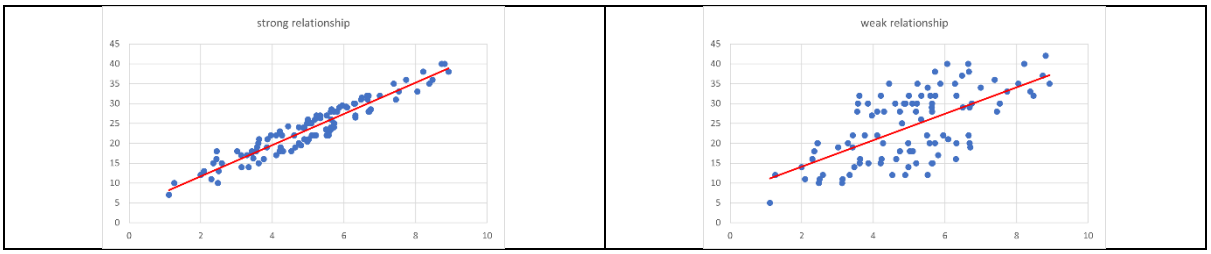
Negative relationship indicates that an increase of the explanatory variable is generating a decrease of response variable.

Some of the frequent **shape** of data are

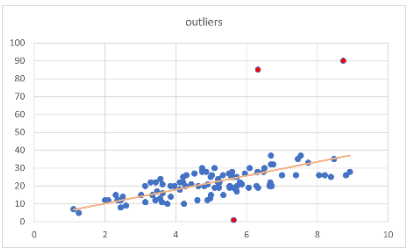




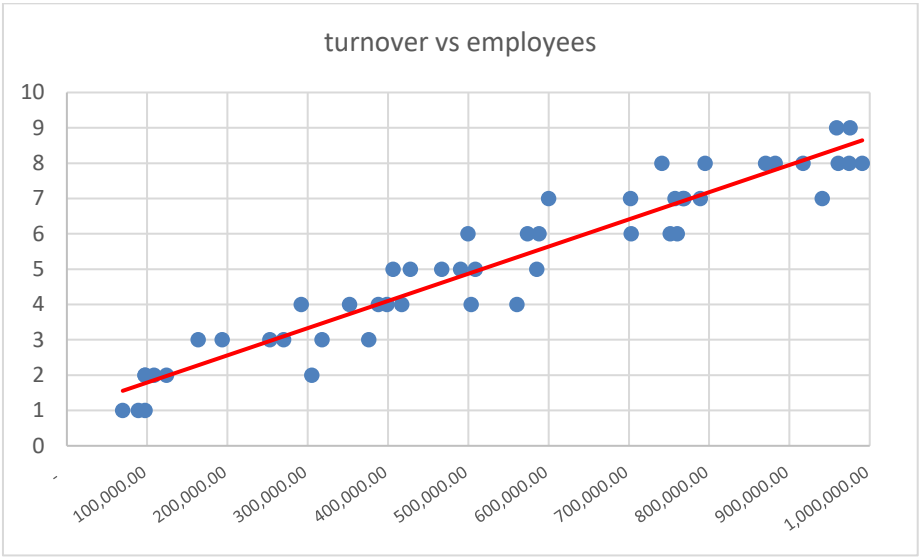
Strength of the relationship is determined by how close data follow the form of the relationship, being possible two scenarios



Deviations from the pattern (outliers) indicate points that are far away from the trend, altering statistics. The following chart is presenting some outliers marked with red



Scatterplot for turnover and employees is



is indicating a linear, positive, and strong relationship. It means that an increase of turnover will determine an increase of employees.

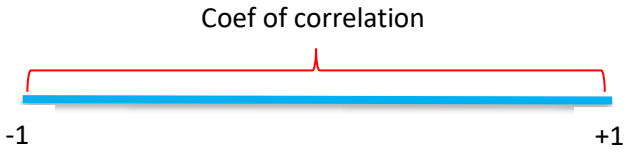
In case of a linear relationship, direction and strength might be assessed by calculating and evaluating coefficient of correlation.

To calculate the coef of correlation, the following formula implemented in Excel might be used

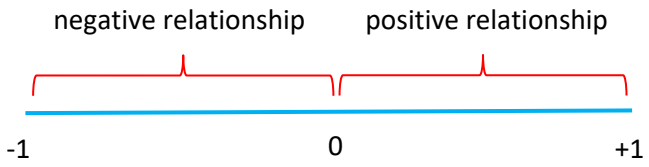
=CORREL (array1, array2)

Usually array 1 refers to explanatory data, while array to refers to response data.

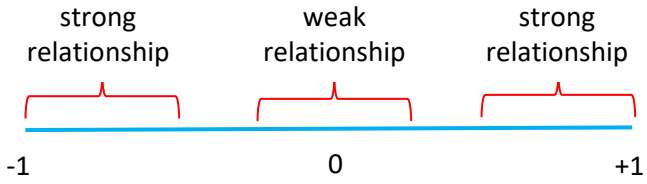
Coef of correlation might have a value between -1 and +1.



A negative value of the coefficient is indicating a negative linear relationship between explanatory and response variables, while a positive value is indicating a positive linear relationship.



Values close to 0 (zero) indicate a weak relationship and values close to -1 or +1 indicate a strong relationship.



In our example the correlation coefficient has a value of 0.96, indicating a strong positive linear relationship between turnover and employees, as it was also suggested by the scatterplot.

Subjective evaluation (weak or strong) for the strength of a linear relationship is not accurate. A numeric measure for the strength of a linear relationship is offered by the coef of determination, calculated as the square of coef of correlation and expressed as percentage.

Coef of determination is expressing *the % from the variation of response variable generated by the variation of explanatory variable.*

In our example coef of determination has a value of 92%, indicating that 92% from the variation of employees is determined by the variation of turnover.